Artificial Intelligence In Medicine 104 (2020) 101815

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Efficient treatment of outliers and class imbalance for diabetes prediction

Nonso Nnamoko, Ioannis Korkontzelos*

Department of Computer Science, Edge Hill University, Ormskirk, United Kingdom

ARTICLE INFO

Keywords:
Outlier detection
Imbalanced data
Machine learning
Data preprocessing
Oversampling
SMOTE

ABSTRACT

Learning from outliers and imbalanced data remains one of the major difficulties for machine learning classifiers. Among the numerous techniques dedicated to tackle this problem, data preprocessing solutions are known to be efficient and easy to implement. In this paper, we propose a selective data preprocessing approach that embeds knowledge of the outlier instances into artificially generated subset to achieve an even distribution. The Synthetic Minority Oversampling TEchnique (SMOTE) was used to balance the training data by introducing artificial minority instances. However, this was not before the outliers were identified and oversampled (irrespective of class). The aim is to balance the training dataset while controlling the effect of outliers. The experiments prove that such selective oversampling empowers SMOTE, ultimately leading to improved classification performance.

1. Introduction

Despite many years of research into machine learning, classification of imbalanced data is still among the major difficulties in the field. The standard learning algorithm assumes that classes within the training dataset are roughly balanced. Also, learning performance metrics often assume equal importance of classes within the dataset. Unfortunately, balanced datasets are rare in real life scenarios and the under-represented class usually has higher misclassification costs [1]. For example, consider the binary classification of the United Kingdom (UK) population as having diabetes or not. Current estimates show that 4.6% of the populace has diabetes [2], which leaves 95.4% non-diabetes cases. A prediction model that classifies all the majority class correctly and all the minority class wrongly would give a very high but misleading Accuracy of 95.4%. The cost of misclassifying those with diabetes may lead to grave consequences.

Classifying outliers is another key issue in machine learning. This problem occurs because data samples rarely follow a clear pattern. In particular, some data samples may share very dissimilar characteristics to others belonging to the same class, and thus become far removed from the mass of data in that class. In medical datasets, such samples may indicate individuals or groups that behave very differently from the majority within the same class. For example, in a binary class task involving a *healthy* and an *unhealthy* group, a heavyweight boxer categorised as *healthy* may be far removed from the mass of observations in this group. This is because the sample is likely to exhibit characteristics that are commonly associated with the *unhealthy* group such as high body mass index (BMI). This dynamic has the potential to disturb the learning mechanism of an algorithm, ultimately leading to misclassification.

Various techniques have been proposed for dealing with outliers and imbalanced datasets, which may be grouped into two broad approaches, namely algorithm level and data level techniques. The former aims at modifying a learning algorithm to cope with the dataset [3] and are known to have relatively high computational cost [4]. The latter is classifier-independent and relatively easy to apply because it focuses on data preprocessing techniques [5]. For example, to deal with outliers, some researchers identify and remove them completely [6], whereas others control the number of outliers to remove [7]. In the same way, some researchers deal with class imbalance by either undersampling the majority class or oversampling the minority class [4].

In this paper, we propose a two-step data preprocessing approach to manage outlier instances and class imbalance. We used the Pima Indians Diabetes dataset[1] obtained from the public UCI data repository [8]. The dataset consists of 768 samples of which 500 tested negative and 268 tested positive. In the first step, we identified the outliers using the Interquartile Range (IQR) algorithm [9] and subsequently oversampled them with replacement [10]. In the second step, we applied the Synthetic Minority Oversampling TEchnique (SMOTE) [11] to obtain a balanced dataset.

* Corresponding author.
E-mail addresses: nnamokon@edgehill.ac.uk (N. Nnamoko), Yannis.Korkontzelos@edgehill.ac.uk (I. Korkontzelos).

[1] Dataset has been removed from UCI repository due to permission restrictions (see archive.ics.uci.edu/ml//datasets/Pima+Indians+Diabetes. However, it is available at kaggle.com/uciml/pima-indians-diabetes-database.)

https://doi.org/10.1016/j.artmed.2020.101815
Received 3 December 2018; Received in revised form 31 January 2020; Accepted 4 February 2020

This research is motivated by the difficulty in identifying individuals at increased risk of developing diabetes. According to Diabetes UK [12], the average diabetes patient has had the condition 9–12 years before it is identified and almost one in 70 people in the UK are living with undiagnosed diabetes. This is mainly because of the reactive management system where a diagnostic test is only prescribed when a patient presents with known complications associated with diabetes [13]. More so, biomedical research studies in this area are mostly retrospective attempts to estimate/project the possible number of undiagnosed cases [14–19]. Therefore, proactive approaches that leads to early identification of individuals at risk of developing the condition is vital so that prevention strategies can be initiated [20–22].

Machine learning has the potential to learn from previous observations such that the resultant model can be used to make proactive decisions on new previously unseen instances. Thus, experiments presented in this paper are based on machine learning classification performed on a diabetes medical health examination dataset. The task is to train learning algorithms with the dataset such that they can predict diabetes onset. Four learning algorithms were considered, namely a Support Vector Machine with a Radial Basis Function kernel (SVM-RBF) classifier [23], C4.5 decision tree [24], Naïve Bayes [25] and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [26]. Their performance is evaluated when parameters such as the percentage of outliers and the minority class are altered on the training data. Thus, rather than eliminate the outliers and/or oversample the minority class only, we initially identified and oversampled the outliers (irrespective of class) before applying SMOTE to balance the dataset. We used SMOTE as it introduces artificial instances on the basis of minority class neighbourhood distribution. The main contributions of this paper include:

- The combination of data preprocessing techniques applied to improve performance. First we oversampled (with replacement) specific data instances (outliers) irrespective of class to alleviate classification bias towards the normal instances. This helped to expose the individual properties embedded in outlier instances to the SMOTE algorithm.
- The classification improvement achieved as a direct result of the data preprocessing combination. The proposed approach led to Accuracy improvements in comparison to similar studies that used the same dataset within the literature.

Evaluation is based on traditional classification metrics, including Accuracy, Precision, Recall, F-score and Cohen's Kappa.[2] We also used McNemar's test to gauge the significance of any improvement made. As *baseline*, we trained AdaBoostM1 [28] and a Random Forest [29] on the *original* dataset. Both classifiers are known to perform well with imbalanced class distribution [28]. Their results are used to measure the performance of our approach when applied to the other four classifiers considered in this study. Furthermore, we examined the validity of our approach in other domains by repeating the experiments with two datasets that exhibits similar characteristics to the Pima diabetes dataset (i.e., large class imbalance and presence of outliers), but unrelated to healthcare.

The rest of the paper is organised as follows. A concise background and related work is provided in Section 2, followed by detailed explanation of the experimental method in Section 3. The results are presented and analysed in Section 4 before concluding Section.

2. Background and related work

Approximately 3.3M adults (4.6% of the UK populace) are now living with diabetes [2] and the number is projected to reach 5 million by 2025 [30]. This figure does not take into account the 549,000 adults estimated to have undiagnosed diabetes [30]. Therefore, it is not surprising that machine learning classifiers are increasingly used to help clinical decision making about diabetes [31]. Typically, machine learning algorithms are used to learn from a sample of observed cases to yield diagnostic or prognostic models that can diagnose or predict new cases. Such learned models might be used to guide physicians' decisions, and are sometimes shown to outperform experts' prediction [32].

The dataset used in this study has also been utilised in a number of other studies [33–39]. These studies approached the classification task differently and achieved varied results. For example, Ramezani et al. [39] proposed a hybrid classifier named Logistic Adaptive Network Based Fuzzy Inference System (LANFIS). LANFIS is a combination of logistic regression and adaptive network-based fuzzy inference system. Basically, LANFIS does not use insignificant attributes during classification and is capable of handling samples with missing values, which is common with the Pima diabetes dataset. Using 3-fold cross validation, LANFIS achieved 88.05% Accuracy on the dataset.

Polat et al. [37] proposed a cascade learning system based on Generalised Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM). The system consists of two stages where GDA is used as a preprocessing tool to discriminate between healthy and diabetes samples; and LS-SVM is used for classification. Applying 10-fold cross validation on the Pima diabetes dataset, the system achieved 82.05% Accuracy which is 3.84% more that the Accuracy obtained with LS-SVM alone (78.21%).

Carpenter and Markuzon [33] examined the Pima diabetes dataset using a neural network classifier called ARTMAP-IC; an extension of ARTMAP neural network [40–42]. Their goal was to solve classification problems where identical input data samples belongs to different classes, which is common in the Pima diabetes data. Using hold-out validation of 576 training and 192 testing samples, ARTMAP-IC produced 81.00% Accuracy.

In a comparative study Kayer and Yildirim [35] also applied hold-out validation of 576 training and 192 testing on Pima diabetes data and compared results with ARTMAP-IC and similar studies. They performed seven experiments with three neural network structures, i.e., five based on Multilayer Perceptron (MLP), and one each for Radial Basis Function (RBF) and General Regression Neural Network (GRNN). Their Accuracy results ranges from 76.56% to 80.21%, which is close to the 81.00% obtained by ARTMAP-IC. However, the highest Accuracy result obtained with GRNN (80.21%) was better than the other 12 studies considered. This includes the Evolving Self-Organising Maps (ESOM) [34] which was better than six state-of-the-art methods with 78.4% Accuracy.

Temurtas et al. [38] used 10-fold cross validation on the Pima diabetes data. However, the authors found that hold-out validation of 576 training and 192 testing produced better Accuracy of 82.37% when applied to a neural network classifiers, i.e., Multilayer Neural Network (MLNN) trained with Levenberg-Marquardt (LM) classifier. They also trained a Probabilistic Neural Network (PNN) with both validation methods but the results were lower.

In fact, our literature search revealed over 70 studies exploiting the Pima diabetes dataset, of which 60 were reported by Winiarski [36] with accuracies ranging between 59.5% and 77.7%. It was not clear how Winiarski [36] validated the models, i.e., hold-out or cross validation.

It is difficult to generalise the reasons behind the varied results obtained in previous studies but data composition, class distribution and certainly the base learning algorithm play important roles. For example, all the studies discussed earlier relied on the base algorithm(s) to produce good performance results. Ideally, these algorithms should extract only the useful patterns from training data and disregard spurious patterns. Unfortunately, training data is usually far from perfect as is the case with our experimental dataset. Some of the

[2] Cohen's Kappa measures the agreement between two raters who each classify *N* items into C mutually exclusive categories [27].

imperfections in the Pima diabetes data include outliers, class imbalance, small sample size and missing values.

Data preprocessing is an alternative method commonly used to alleviate imperfections such as imbalance class ratio. Most often this is achieved by either under or oversampling as required [4]. For example, the minority class can be oversampled with or without replacement. Sampling with replacement means that an object is replaced each time it is drawn from a pool of objects and can be re-drawn [10]. This method is often used when the oversampling requirement is more than the quantity available within the pool. However, more sophisticated approaches exist such as SMOTE [11], which generates new synthetic examples by using knowledge about neighbours surrounding a given object within the pool, as discussed further in Section 2.1.

The original SMOTE has been extensively studied with several improvements proposed. For example, Chawla et al. [43] proposed SMOTEBoost by combining the original SMOTE algorithm with a boosting procedure. The Borderline-SMOTE [44] oversamples minority data samples in unsafe regions. The Safe-Level-SMOTE [45] took the opposite approach by focusing on the safest data samples. LN-SMOTE [46] exploits the local information about the neighbourhoods of oversampled data samples. MWMOTE [47] extended the SMOTE algorithm by modifying the synthetic data generation process to use a clustering approach.

While over-sampling is the dominant approach, several studies have focused on under-sampling. Among them, the neighbourhood cleaning rule [48] that removed minority class instances overlapping heavily with the majority class. Liu et al. [49] combined under-sampling with ensemble classifiers to improve performance. García and Herrera [50] combined under-sampling with evolutionary algorithms. The study was later extended by Galar et al. [51] who added a boosting algorithm. Koziarski and Wozniak [52] combined under-sampling and over-sampling to improve classification performance. They initially cleaned the neighbourhoods of minority samples by removing objects from the majority class. This was aimed at simplifying the task of classifying samples from the minority class. They then selectively oversampled the minority class instances by generating synthetic samples closer to the minority instances in the least safe zone, i.e., where the majority class is dominant.

In fact, we found two studies [53,54] that specifically sought to tackle class imbalance in the Pima diabetes data through some form of under-sampling. Raghuwanshi and Shukla [54] proposed an Under-bagging Based Kernelised Extreme Learning Machine (UBKELM) for dealing with class imbalance. The algorithm creates several balanced training subsets by random under-sampling of the majority class samples. Then a kernalised Extreme Learning Machine (ELM) is used as the component classifier to make ensembles. This was tested with the Pima diabetes data which produced 75.84% G-mean[3] and 80.55% AUC.[4]

Nanni et al. [53] also used the Pima diabetes data to test two ensemble of ensembles methods designed to tackle class imbalance. Their approach is based on under-sampling like Raghuwanshi and Shukla [54], except it is not done randomly. One of the methods, called EasyEnsemble, samples the majority class into several independent subsets that are used to train separate classifiers. These outputs are combined to produce the classification decision. The second method called BalanceCascade is focused on training patterns that are hard to classify and trained models are used to guide the sampling process for succeeding classifiers. Their best results for the Pima diabetes data are 84.18% AUC, 69.17% F-score[5] and 75.77% G-mean.

Jegierski and Saganowski [55] recently proposed an 'outside the box' solution to class imbalance in which external but similar data was used to enrich samples of the minority class. They used various datasets to test three data enrichment options namely Random Enrichment (RanE), Semi-greedy Enrichment (SemE) and Supervised Enrichment (SupE). RanE simply selects samples at random from the external dataset and adds them to the minority class. SemE iteratively selects/validates the samples from the external dataset that would increase classification performance. SupE only selects borderline samples from the external dataset to help define boundaries between classes. Their approach performed better than nine well known methods for mitigating class imbalance, including four versions of SMOTE.

It is, however, important to note that class disproportion in data usually does not pose a problem by itself. Local characteristics of the minority class are equally as important [4]. According to Stefanowski et al. [56], class imbalance only affects minority class recognition when combined with other data difficulty factors such as outliers, overlapping class, etc. Therefore, such factors (outliers in our case) must be considered when exploring new ways of dealing with imbalanced data. Napierała et al. [57] measured the impact of noisy and borderline data samples from the minority class on classification performance. They found that the degradation in performance of a classifier is strongly affected by the number of borderline data samples. Skryjomski and Krawczyk [4] also proposed a method of improving classification performance by oversampling borderline data samples from the minority class with SMOTE.

The approach proposed in this paper is not one-sided (i.e., focused on minority class only) but focuses on the entire dataset (irrespective of class). We identified outliers from the *original* dataset and then oversample the instances with replacement. The aim of this step is to increase the number of these rare cases within the dataset so that when SMOTE is applied to obtain class balance, more synthetic instances would be generated near the neighbourhood of the outliers. In effect, our approach exposes the learning algorithm to more rare cases, which may be difficult to learn. No other study within the literature was found to have combined these data preprocessing techniques in the same way. Before detailing the setup of the experiments, the data preprocessing techniques, SMOTE and IQR, used in the experiments are briefly discussed.

2.1. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is an oversampling technique introduced by Chawla et al. [11]. As opposed to other methods that oversample instances randomly by duplication, SMOTE creates new artificial instances using knowledge about neighbours that surround each sample in the minority class. The pseudocode in Algorithm 1 describes the method.

Algorithm 1. *SMOTE* algorithm [4]

$D_{smoted} \leftarrow [\,]$;
for $i \leftarrow 1$ **to** nrow $(D_{minority})$ **do**
  $nn \leftarrow k$-NN $(D_i, D_{minority}, k)$;
  $N_i \leftarrow \lfloor N_{percent}/100 \rfloor$;
  while $N_i! = 0$ **do**
    $neighbour \leftarrow$ SelectRandom (nn);
    $gap \leftarrow$ RangeRandom$(0, 1)$;
    $diff \leftarrow neighbour - D_i$;
    $synth \leftarrow D_i + gap \times diff$;
    $D_{smoted} \leftarrow$ append$(D_{smoted}, synth)$;
    $N_i \leftarrow N_i - 1$;
  end
end

SMOTE uses k-nearest neighbour (k-NN) algorithm to find the k

[3] G-mean stands for geometric mean and is defined as the root of the product of class-wise sensitivity, i.e., true positive rate and true negative rate

[4] AUC means area under the receiver operating characteristic (ROC) curve. ROC curve is a plot of the sensitivity vs. false positive rate (1 minus specificity).

[5] F-score is the harmonic mean of Recall and Precision

nearest neighbours of a given minority data instance data from the neighbourhood. (Note: k is an integer value provided as an input). In the pseudocode, $D_{minority}$ is the number of minority class instances and $N_{percent}$ is the percentage of instances to be generated by SMOTE. Neighbourhood distance can be calculated with various metrics, but for the experiment reported in this paper, Euclidean distance is used. The Euclidean distance between two points x_i and x_j is the length of the line segment connecting them ($x_i x_j$) [58]. k neighbours (5 in our case) are identified for each data item. To generate an artificial instance, one of these k neighbours of an original minority instance is chosen at random and used for further processing. The number of synthetic instances formed per original instance is determined by $N_{percent}$, which is supplied as an input to the SMOTE algorithm. Each new instance is created by adding to the features of the original minority instance (D_i) the differences (*diff*) between the corresponding features of the chosen neighbour instance and the original instance; multiplied by a random number (*gap*) between 0 and 1. This helps to determine the final position of the generated instance, which may be in the same position as the original minority instance, the randomly selected neighbour or anywhere between the two locations. By so doing, the diversity of the generated instance is increased thereby allowing for better exploitation of the decision space.

2.2. The Interquartile Range (IQR) algorithm

IQR is a data preprocessing technique used to detect outliers and extreme values. It measures dispersion by dividing a rank-ordered dataset into four equal parts, called quartiles [9]. The values that divide each part are denoted by Q1, Q2, and Q3, where Q1 and Q3 are the middle value in the first and second half of the rank-ordered dataset respectively; and Q2 is the median value in the entire set. IQR is then equal to Q3 minus Q1. Outliers here are data instances that fall below Q1 − 1.5 IQR or above Q3 + 1.5 IQR.

In the boxplot in Fig. 1, the highest and lowest occurring values within this limit are indicated by whiskers of the box and any outliers as individual points. Q1, Q2 and Q3 are 7, 8.5 and 9 respectively. The IQR = Q3 − Q1 = 2. The lower whisker = Q1 − 1.5 × IQR = 7 − 3 = 4. The upper whisker = Q3 + 1.5 × IQR = 9 + 3 = 12. Data points 0.5 and 3.5 are outliers, perhaps of different classes.

3. Method

In this section, we present our approach to reduce classification bias towards a class in the training data, whilst acknowledging the presence of outlier instances. The approach involves multiple preprocessing of the training data as illustrated in Fig. 2. Firstly, we search for outliers in the *original* training data using IQR algorithm. The outlier instances are then oversampled with replacement and subsequently added back to the *original* data. The oversampling percentage is chosen arbitrarily depending on the number of outliers in the data, with the ultimate aim to massively increase their presence in the data. As the process may lead to class imbalance, we introduce SMOTE to even the class distribution before classification. A detailed description of the data is presented in Section 3.1, followed by the experimental setup in Section 3.2.

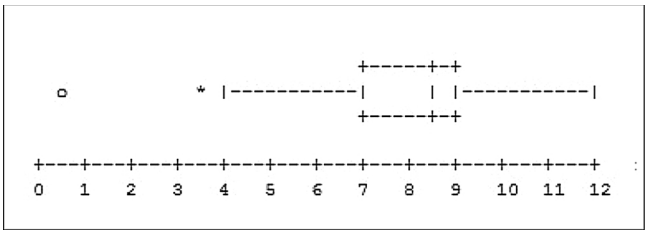

Fig. 1. Sample box plot showing outliers.

3.1. Datasets

For experimentation purposes, a diabetes dataset was obtained from the public UCI data repository [8]. It consists of 768 females of Pima Indian origin aged 21 or above who took part in a national health check program aimed at diagnosing diabetes. There are 500 negative and 268 positive instances. 9 features were obtained for each individual, including the class variable as shown in Table 1.

Variations of the *original* data were generated and used to train the classification algorithms considered in this study. These variations are shown in Table 2. The *SMOTEd* data was obtained by oversampling the minority class in the *original* data using SMOTE. The oversampling ratio was set to 90% (n = 241) so that the classes are approximately balanced, i.e., 509 positive and 500 negative instances. The *IQRd + SMOTEd* dataset was generated using the method described in Section 3. We searched for outliers in the *original* data using the IQR algorithm, which identified 49 outlier instances. These instances were oversampled with replacement by 500%, resulting in 245 outliers; and subsequently added back to the *original* data, i.e., (768 − 49) + 245 = 964 instances. We oversampled with replacement due to the small number of outliers that would otherwise not allow generation of 196 additional outlier instances. The oversampling percentage, i.e., 500% was chosen arbitrarily with the ultimate aim to massively increase the number of outlier instances. This process did not lead to a balanced dataset, as outliers were oversampled irrespective of class. Thus, we oversampled the minority class of the new data by 50% (n = 193) using SMOTE so that the classes are evenly distributed, i.e., 579 positive and 578 negative instances.

3.2. Experiment setup

The three data variations described in Section 3.1 were used to train four classifiers, namely: Naïve Bayes, SVM-RBF, C4.5 decision tree and RIPPER. These classifiers were selected because of their popularity in the field of machine learning. To generate a *baseline* against which our approach is measured, we also trained AdaBoostM1 [28] and Random Forest [29] classifiers using the *original* data, i.e., without preprocessing. We choose these classifiers because they are known to mitigate against skewed class distribution in training data [28]. AdaBoostM1 works by repeatedly running an *underlying* learning algorithm on various distributions over the training dataset, and then combining their outputs into a single composite classifier. We applied Decision Stump [59] as the *underlying* algorithm for AdaBoostM1 in this paper. A Decision Stump is simply a one-level Decision Tree model that makes prediction based on the value of a single input feature. Random Forest however, is based on a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [29].

Five traditional performance metrics were considered, to evaluate performance, including Accuracy, Precision, Recall, F-score and Kappa [27]. With the exception of Kappa, these metrics are interpreted on a scale of 0 (lowest) to 1 (highest). Kappa, is interpreted on a scale of −1 (lowest) to 1 (highest). To estimate the significance of improvement as a result of the proposed approach, McNemar's test [60] was used. This is a non-parametric test on a 2 × 2 classification table, shown in Table 3, to measure the difference between paired proportions.

N_{ff} denotes the number of times both classifiers failed to classify instances correctly and N_{ss} denotes success for both classifiers. These two values do not give much information about the classifiers' performances as they do not indicate how their performances differ. The other two parameters, N_{sf} and N_{fs}, reflect cases where one of the classifiers failed and the other succeeded, indicating performance discrepancies.

Multiple independent evaluation using stratified k-fold (k = 10) cross-validation was performed to gauge the performance of our approach. In stratified k-fold cross-validation, the training data is

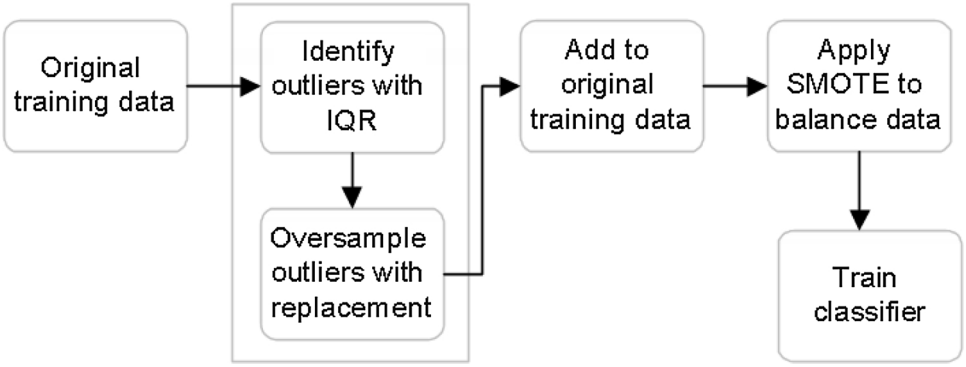

Fig. 2. High level diagram of the proposed method.

Table 1
Experimental data features.

Feature #	Description
1	Number of times pregnant
2	Plasma glucose concentration in a 2 h oral glucose tolerance test
3	Diastolic blood pressure (mmHg)
4	Triceps skin fold thickness (mm)
5	2-h serum insulin (μIU/ml)
6	Body mass index (weight in kg/(height in m^2)
7	Diabetes pedigree function
8	Age (years)
9	Class (0, 1)

Table 2
Characteristics of the original and preprocessed datasets.

Dataset	Positive	Negative	Total
Original	268	500	768
SMOTEd	509	500	1009
IQRd + SMOTEd	579	578	1157

Table 3
Simple *McNemar's* table showing result of two classifiers.

	Classifier B failed	Classifier B succeeded
Classifier A failed	N_{ff}	N_{fs}
Classifier A succeeded	N_{sf}	N_{ss}

randomly partitioned into 10 equal size subsets, taking the class distribution into account. During training, one of the k subsets is retained as the validation data, and the remaining $k-1$ subsets are used as training data. The process is repeated k times, with each of the k subsets used exactly once as the validation data. The k results from the folds are then combined to produce a single result.

To ensure that the results are unbiased, while at the same time allow for cross comparison between different classifiers, we used the exact folds generated from the *original* data, to evaluate all the classifier models trained with the *preprocessed* versions of the data. Basically, stratified k-fold cross validation was applied to the *original* data, to generate 10 folds. Then, the *original* data was *preprocessed* to generate the *SMOTEd* and the *IQRd + SMOTEd* version of the data. In each cross validation iteration, we considered as the test set one fold of the *original* dataset. All the instances of this fold was removed from the *preprocessed* version of the data, and the remaining instances as considered training data for that iteration. This means that each training fold in *SMOTEd* or *IQRd + SMOTEd* datasets, consists of the full *preprocessed* dataset minus the instances of the test fold. The experimental process is illustrated in Fig. 3.

4. Result analysis

In this section, we present and analyse the performance of the four classifiers when trained with the *original* and *preprocessed* datasets described in Section 3.1; particularly *IQRd + SMOTEd*. We also repeated the experiments with other datasets that are unrelated to diabetes. This is to examine if the proposed method extends to other datasets and domains. For clarity, results with the Pima diabetes and *other unrelated* datasets are analysed in separate Sections 4.1 and 4.2 respectively. Each Section also presents the corresponding *baseline* results obtained with AdaBoostM1 and Random Forest.

4.1. Results with diabetes dataset

This section presents the results obtained with the Pima diabetes datasets described in Section 3.1. Table 4 shows the evaluation results for each classifier trained with the *original* and *preprocessed* Pima diabetes datasets. For each group of classifier models presented in the table, we use **bold typeface** to indicate the best on all the performance metrics. It is clearly evident that the models trained with *IQRd + SMOTEd* dataset consistently produced the best Accuracy. Of the four classifiers considered, C4.5 produced the best Accuracy while Naïve Bayes produced the least. However, it is important to note that even the least performing Naïve Bayes model produced better Accuracy than the *baseline*. In terms of Kappa, the proposed method using *IQRd + SMOTEd* also led to better results than the *baseline* in all but SVM-RBF classifier where performance is lower.

Generally, Naïve Bayes and SVM-RBF produced mixed results in all the performance metrics and they clearly did not respond well to the selective data preprocessing method applied. For example, Naïve Bayes trained with *SMOTEd* produced slightly better Precision than *IQRd + SMOTEd* by 0.001%. Similarly, the SVM-RBF model is 0.60% more precise when trained with *SMOTEd* dataset than with *IQRd + SMOTEd*. However, it is fair to say that these differences are marginal and unlikely to be significant.

While mixed results were recorded for Naïve Bayes and SVM-RBF, the other two classifiers, i.e., RIPPER and C4.5 responded well to *IQRd + SMOTEd* data, as reflected in all the performance metrics. In fact, McNemar's test conducted on the models show that both classifiers trained with *IQRd + SMOTEd* data produced statistically significant improvement when compared to their performance with the other training data versions. This result can be seen clearly in Table 5, which presents the results of *IQRd + SMOTEd* trained models vs. *original* and *SMOTEd* for all the classifiers. Statistically significant differences between two models are indicated with '★' sign. The level of improvement can be seen clearly in the prediction success (N_{fs}) and failure N_{sf}) columns of Table 5, which translates into the value presented in the *diff* column. For example, taking the counts for success and failure into account, models that led to significant difference succeeded in predicting the true class more often (N_{fs}) than predicting wrongly (N_{sf}). The

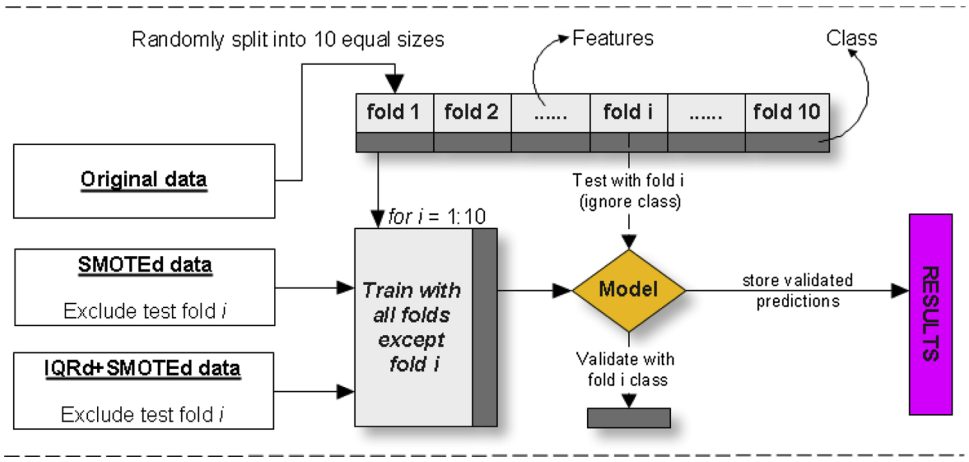

Fig. 3. High level diagram of the experiment setup.

differences between (N_{fs}) and (N_{sf}) were minimal with Naïve Bayes and SVM-RBF which explains the insignificant improvement. On the other hand, the differences were larger with RIPPER and C4.5, ultimately leading to significant improvement.

As observed earlier in Table 4, C4.5 classifier trained with *IQRd + SMOTEd* data produced the best results overall. Its Accuracy improvement over the other classifiers ranges from 5.9% to 12.5%. To experimentally demonstrate the significance of this improvement, we conducted a McNemar's test to compare its predictions with the best performing models of the other classifiers, including the *baseline*: AdaBoostM1 and Random Forest. The result is shown in Table 6, which confirms that the differences are indeed significant. The proposed method proves to make significant improvement in classification performance when applied to a classifier that responds well to the *preprocessed* dataset, i.e., C4.5 in this case.

As noted in Section 2, we found over 70 published studies that utilised the Pima diabetes dataset, so we compared their results to ours in Table 7. Clearly our selective data preprocessing approach applied to C4.5 classifier produced better Accuracy, than the past studies with improvements ranging from 1.45% to 30%. It must be noted that methods (i.e., classification approach) and validation approach (i.e., training and test data split) differ between these studies and this may affect the reported results. For example, Temurtas et al. [38] applied both 10-fold cross validation and a 576:192 data split to the same classifier and found the latter more favourable as shown in Table 7. Of all the past studies presented in Table 7, Ramezani et al. [39] produced the best Accuracy (88.05%), by applying 3-fold cross validation on LANFIS. Quite possibly, their result would be different if they used a different validation approach. Our view is that these results are comparable to an extent, irrespective of the method or validation approach used. This is because all the studies used the same data for experiments and the common goal is to improve results with the proposed method whilst ensuring that the test dataset is not exposed to the classifier during training. Moreover, given the modest size of the Pima diabetes data, 10-fold cross validation used in our experiment seems more appropriate because testing is based on all samples rather than a small subset of the data. Even when we compare with only the past studies that used 10-fold cross validation [37,38,34], our approach produced better Accuracy, with improvements ranging from 7.13% to 11.1%.

Some of the previous studies discussed in Section 2 did not measure performance based on Accuracy, hence not presented in Table 7. In particular, Nanni et al. [53] reported performance in terms of F-score, G-mean and AUC while Raghuwanshi and Shukla [54] reported only G-mean and AUC. Both studies specifically aimed to mitigate class imbalance in training data and they tested their methods with the Pima diabetes data. As C4.5 (*IQRd + SMOTEd* is our best performing model on the Pima diabetes data, we calculated F-score, G-mean and AUC, and compared to the best results from Nanni et al. [53] and Raghuwanshi and Shukla [54]. As shown in Table 8, our approach performed better in all the three metrics considered.

Table 4
Classifier performance with *original* and *preprocessed* versions of the Pima diabetes dataset.

Classifier/data		Accuracy	Precision	Recall	F-Score	Kappa
Baseline	AdaBoostM1	0.746	0.740	0.746	0.741	0.638
	Random Forest	**0.755**	**0.749**	**0.755**	**0.750**	**0.650**
Naïve Bayes	Original	0.760	0.758	0.760	0.760	0.650
	SMOTEd	0.762	**0.770**	0.762	0.764	0.611
	IQRd + SMOTEd	**0.770**	0.769	**0.769**	**0.768**	**0.653**
SVM-RBF	Original	0.760	0.758	0.760	0.755	**0.660**
	SMOTEd	0.768	**0.792**	0.768	0.773	0.585
	IQRd + SMOTEd	**0.777**	0.786	**0.777**	**0.779**	0.641
RIPPER	Original	0.772	0.772	0.772	0.766	0.676
	SMOTEd	0.776	0.808	0.776	0.781	0.582
	IQRd + SMOTEd	**0.836**	**0.842**	**0.836**	**0.836**	**0.743**
C4.5	Original	0.747	0.753	0.747	0.744	0.621
	SMOTEd	0.810	0.835	0.810	0.814	0.653
	IQRd + SMOTEd	**0.895**	**0.900**	**0.894**	**0.895**	**0.835**

Table 5
Mc Nemar's test showing performance differences between the Pima diabetes data versions. Each dataset in the second column is compared against *IQRd + SMOTEd*.

Classifier/data		N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
NB	Original	162	22	15	569	0.91	[−0.64, 2.46]	0.3240
	SMOTEd	157	26	20	565	0.78	[−0.95, 2.51]	0.4614
SVM	Original	136	48	55	549	1.69	[−0.63, 4.01]	0.1875
	SMOTEd	125	53	46	544	0.91	[−1.63, 3.45]	0.5467
RIPPER	Original	81	94	45	548	6.38	[3.41, 9.36]	< 0.0001★
	SMOTEd	78	94	48	548	5.99	[2.98, 9.00]	0.0001★
C4.5	Original	63	131	18	556	14.71	[11.78, 17.65]	< 0.0001★
	SMOTEd	59	87	22	600	8.46	[5.87, 11.06]	< 0.0001★

N_{ff}: both models failed — N_{fs}: *IQRd + SMOTEd* trained model succeeded and the other model failed
N_{ss}: both models succeeded — N_{sf}: *IQRd + SMOTEd* trained model failed and the other model succeeded

Table 6
IQRd + SMOTEd trained *C4.5* vs. best of other classifiers including baseline models.

Classifier	N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
AdaBoostM1 vs *C4.5*	64	131	17	556	14.84	[11.92, 17.77]	< 0.0001★
Random Forest vs *C4.5*	35	153	46	534	13.93	[10.47, 17.39]	< 0.0001★
NB vs *C4.5*	63	114	18	573	12.50	[9.70, 15.30]	< 0.0001★
SVM-RBF vs *C4.5*	69	102	12	585	11.72	[9.12, 14.31]	< 0.0001★
RIPPER vs *C4.5*	53	73	28	614	5.86	[3.33, 8.39]	< 0.0001★

N_{ff}: both models failed — N_{fs}: *C4.5* succeeded and the other classifier failed
N_{ss}: both models succeeded — N_{sf}: *C4.5* failed and the other classifier succeeded

4.2. Results with other Datasets

This Section presents the results obtained by applying our method described in Section 3 to other datasets that are unrelated to diabetes. We believe this is necessary because the modest size of the Pima diabetes data did not allow for an expansive validation process, e.g., having a hold-out set for testing purposes. The ideal scenario would be to test on a similar but external dataset but we could not find another dataset with similar features and class labels.

Recall from Section 2 that Jegierski and Saganowski [55] proposed three strategies, i.e., Random Enrichment (RanE), Semi-greedy Enrichment (SemE) and Supervised Enrichment (SupE), for class imbalance that uses external but similar data to enrich samples of the minority class. When tested with Random Forest [29], on a *breast cancer* dataset,[6] the strategies were shown to produce better F-score than nine well known methods for mitigating class imbalance, including four versions of SMOTE.

For the purpose of comparison, we replicated our approach with Random Forest on the same *breast cancer* dataset, which contains 569 samples of which 357 are *Benign* and 212 are *Malignant*. Basically, we identified 11 outliers and these were oversampled with replacement by 500% before applying SMOTE to increase the minority class by 35%. For testing, we applied 10-fold cross validation on the *original* data and computed F-score average just like Jegierski and Saganowski [55]. As shown in Table 9, our approach is comparable to SupE (without SemE) but better than RanE, SemE and SupE (with SemE).

To determine if our approach applies to domains outside healthcare, we also replicated our experiments (described in Section 3.2) with two datasets that are unrelated to healthcare. Both datasets, i.e., German Credit[7] and QSAR Biodegradation[8] were obtained from the public UCI data repository [8], and they exhibit similar characteristics to the Pima diabetes data, i.e., heavily imbalanced and containing outliers. Table 10 shows the class distribution for both datasets including their *original* and *preprocessed* versions.

So that the paper remains focused, only the highlights of experimental results obtained with both datasets are discussed within the main body of this paper. Further details of the experiments are presented in Appendix A for the German Credit dataset, and Appendix B for the QSAR Biodegradation dataset.

As expected, the results follow similar pattern to the one obtained with the Pima diabetes dataset. Apart from experiments with Naïve Bayes, which produced mixed results across the performance metrics, all the experiments involving *IQRd + SMOTEd* data produced the best results within their respective classifier groups. In particular, SVM-RBF trained with *IQRd + SMOTEd* produced the best overall performance on both experimental datasets, i.e., German Credit and QSAR Biodegradation. It is also important to note that the performance with *IQRd + SMOTEd* (except experiments with Naïve Bayes) are better than or closely matches the *baseline*. This is shown clearly in the appended Tables A.1 and B.1 .

For each of the classifiers considered in this study, we also tested the significance of improvement between the *IQRd + SMOTEd* trained model and others. Again, significant difference was observed in most cases, particularly when compared to classifiers trained with the *original* data, but not their *SMOTEd* versions as shown in the appended Tables A.2 and B.2 . Although our method did not always lead to significant

[6] Breast cancer dataset is available at archive.ics.uci.edu/ml/datasets/Breast + Cancer + Wisconsin + (Diagnostic).

[7] German credit data is available at archive.ics.uci.edu/ml/datasets/statlog + (german + credit + data).

[8] QSAR Biodegradation data is available at archive.ics.uci.edu/ml/datasets/QSAR + biodegradation.

Table 7
Comparative results with previous studies based on *Accuracy*.

Author/article	Method	Accuracy
Our work	C4.5 (*IQRd + SMOTEd*) **Validation:** 10-fold CV	89.5%
Winiarski [36]	Comparative study with 60 different classifiers **Validation:** undisclosed	Range = 59.5–77.7%
Polat et al. [37]	• Least Square Support Vector Machine (LS-SVM) • Ensemble of Generalised Discriminant Analysis (GDA) and LS-SVM **Validation:** 10-fold CV	• LS-SVM = 78.21% • GDA + LS-SVM = 82%
Kayaer and Yildirim [35]	• General Regression Neural Network (GRNN) • Radial Basis Function (RBF) • Levenberg–Marquardt (LM) • Gradient Descent (GD) • GD with momentum (M) • GD with M and adaptive learning rate (ALR) • BFGS quasi Newton **Validation:** hold-out (576:192)	• GRNN = 80.21% • RBF = 68.23% • LM = 77.08% • GD = 77.60% • GD + M = 76.56% • GD + M + ADL = 77.60% • BFGS = 77.08%
Temurtas et al. [38]	• MLNN + LN (576:192) • PNN (576:192) • MLNN + LN (10-fold CV) • PNN (10-fold CV)	• MLNN + LN = 82.37% • PNN = 78.13% • MLNN + LN = 79.62% • PNN = 78.05%
Carpenter and Markuzon [33]	ARTMAP-Instance Counting (ARTMAP-IC) **Validation:** hold-out (576:192)	ARTMAP-IC = 81%
Deng and Kasabov [34]	Evolving Self Organising Maps (ESOM) **Validation:** 10-fold CV	ESOM = 78.4%
Ramezani et al. [39]	Logistic Adaptive Network-based Fuzzy Inference System (LANFIS) **Validation:** 3-fold CV	LANFIS = 88.05%

Table 8
Comparative results with previous studies based on *G-Mean*, *AUC* and *F-Score*.

Author/article	AUC (%)	G-mean (%)	F-Score (%)
Our work – C4.5 (*IQRd + SMOTEd*)	94.6	88.8	89.5
Nanni et al. [53]	84.2	75.8	69.2
Raghuwanshi and Shukla [54]	81.6	75.8	N/A

Table 9
Comparative results with breast cancer data based on *F-Score*.

Method/author	F-Score (%)
Our work – Random Forest (*IQRd + SMOTEd*)	94.7
RanE [55]	92.3
SemE [55]	93.6
SupE (with SemE) [55]	86.5
SupE (without SemE) [55]	95.1

Table 10
Characteristics of the two datasets used for method validation in other domains.

Data/variation		Positive	Negative	Total
Credit	Original	700	300	1000
	SMOTEd	700	690	1390
	IQRd + SMOTEd	722	723	1445
Biodeg	Original	356	699	1055
	SMOTEd	697	699	1396
	IQRd + SMOTEd	946	948	1894

difference, it certainly produces some form of improvement in all but Naïve Bayes classifier. Such performance provides a base that could be built upon for further improvement. For example, the classifiers that improved performance but not enough to effect significant difference could be optimised internally to improve prediction Accuracy.

We know from Tables A.1 and B.1 that SVM-RBF trained with *IQRd + SMOTEd* produced the best results. To experimentally demonstrate the significance of this improvement, over the best performing models from the other classifiers, including the *baseline* models – AdaBoostM1 and Random Forest; we conducted a McNemar's test to compare their predictions. The results are shown in the appended Tables A.3 and B.3 . Once again, the results are very similar to that obtained with the Pima diabetes data. The performance of SVM-RBF trained with *IQRd + SMOTEd* data led to significant improvement in all but one classifier, i.e., Random Forest in the appended Table B.3. Nevertheless, the result is a clear indication that given the right classifier, models trained with the selective data preprocessing method presented in this study generally responds positively to class imbalance and outliers.

5. Conclusion

The experiments presented in this paper proved our intuition that the selective data preprocessing method proposed in this paper can be used to achieve greater Accuracy than existing work conducted with the Pima diabetes data. To achieve this, we examined the effects of outliers and data imbalance on classification performance. The reason for combining the two factors is because literature evidence suggests that class imbalance is not the only source of learning difficulties posed by training data during classification tasks. We first analysed the entire

data distribution to check for factors other than class imbalance, with potential to cause misclassification. This resulted to outliers which are usually sparse within the training data, thus difficult for classifiers to comprehend. As a solution, we identified outliers with IQR algorithm, enhanced their presence through oversampling and tackled class imbalance with SMOTE. By initially oversampling the outlier instances and subsequently creating instances within their neighbourhood distribution through SMOTE, the proposed method affords the learning algorithms wider visibility of the sparse instances and thus a better learning platform to improve performance.

Experiments with Naïve Bayes, SVM-RBF, C4.5 and RIPPER show that our selective data preprocessing method applied to C4.5 decision tree produced better results than the other three classifiers with 89.5% Accuracy, 90% Precision, 89.4% Recall, 89.5% F-score and 83.5% Kappa. These results are also better than *baseline* experiments conducted with AdaBoostM1 and Random Forest. Indeed, the vast majority of reported experiments in diabetes prediction only enhanced classification Accuracy up to 82% [61], which is 7.5% lower than our results. However, not all available diabetes prediction study in the literature is based on the same dataset applied to our study, so we identified those with the same dataset and compared results. The search revealed a total of 71 studies exploiting the same dataset, with Accuracy results ranging between 59.5% and 88.05%. This clearly shows that our approach increased Accuracy in the range between 1.45% and 30%. By selectively exposing SMOTE to knowledge about outliers (irrespective of class), our method led to improved performance. This is also the case when compared with uniform oversampling of minority class, without outlier selection.

As validation in other domains, we applied our method to two datasets that are unrelated to the medical domain. This time, the SVM-RBF rather than C4.5 produced the best results and performance is consistently better with our data preprocessing method, in all the classifiers except Naïve Bayes. In the future, we plan to develop the data preprocessing method as a standalone tool that identifies and embeds outlier knowledge into the standard version of SMOTE algorithm. This will enable us to compare between our method and other versions of the SMOTE algorithm described in Section 2. We also plan to investigate why Naïve Bayes does not respond well to the method. Perharps this line of research would provide useful patterns that can be used to determine based on a given dataset, how different classifiers would respond to the method.

Conflict of interest

None declared.

Acknowledgement

This research has been carried out as part of the *CROSSMINER Project*, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under *Grant Agreement No. 732223*.

Appendix A. German credit data

The German credit data from the finance domain is used to determine if a person is credit worthy or not. The data contains 1000 samples, each with 20 features that characterises the samples such as *age*, *employment*, etc. The *SMOTEd* version of the data was generated by simply oversampling the minority class by 130% using SMOTE. From the *original* version, we identified 25 outlier instances using IQR algorithm. These instances were oversampled (with replacement) by 200% to generate additional 50 and subsequently added back to the data, i.e., a total of 75 outliers now exist. Finally, the minority class was oversampled by 105% using SMOTE to generate the *IQRd + SMOTEd* version of the data.

The results are presented in Tables A.1–A.3. We use **bold typeface** to indicate best performance among classifier groups, and '★' sign to indicate statistically significant difference.

Table A.1
Classifier performance with *original* and *preprocessed* versions of the *credit* dataset.

Classifier/data		Accuracy	Precision	Recall	F-Score	Kappa
Baseline	AdaBoostM1	0.719	0.688	0.714	0.687	−0.735
	Random Forest	**0.772**	**0.758**	**0.767**	**0.750**	**−0.274**
Naïve Bayes	Original	**0.762**	0.752	**0.759**	0.754	0.059
	SMOTEd	0.743	0.753	0.741	**0.764**	**0.234**
	IQRd + SMOTEd	0.759	**0.762**	0.757	0.759	0.223
SVM-RBF	Original	0.716	0.689	0.708	0.615	−8.793
	SMOTEd	0.939	0.939	0.939	0.939	0.798
	IQRd + SMOTEd	**0.953**	**0.953**	**0.953**	**0.953**	**0.843**
RIPPER	Original	0.735	0.718	0.731	0.721	−0.173
	SMOTEd	0.770	0.770	0.768	0.769	0.236
	IQRd + SMOTEd	**0.772**	**0.775**	**0.770**	**0.772**	**0.272**
C4.5	Original	0.719	0.698	0.715	0.702	−0.784
	SMOTEd	0.819	0.818	0.817	0.817	0.380
	IQRd + SMOTEd	**0.836**	**0.835**	**0.835**	**0.835**	**0.444**

Table A.2
Mc Nemar's test showing performance differences between the *credit* data versions. Each dataset in the second column is compared against *IQRd + SMOTEd*.

Classifier/data		N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
NB	Original	173	65	78	694	−0.30	[−2.56, 1.96]	0.8624
	SMOTEd	97	161	144	598	1.70	[−1.72, 5.12]	0.3596
SVM	Original	16	268	31	685	23.70	[20.65, 26.75]	< 0.0001★
	SMOTEd	31	30	16	923	1.40	[0.073, 2.73]	0.0541
RIPPER	Original	145	120	83	652	3.70	[0.92, 6.48]	0.0113★
	SMOTEd	142	88	86	684	0.20	[−2.39, 2.79]	0.9396
C4.5	Original	109	172	55	664	11.70	[8.84, 14.56]	< 0.0001★
	SMOTEd	126	55	38	781	1.70	[−0.19, 3.59]	0.0966

N_{ff}: both models failed — N_{fs}: *IQRd + SMOTEd* succeeded and the other model failed
N_{ss}: both models succeeded — N_{sf}: *IQRd + SMOTEd* failed and the other model succeeded

Table A.3
IQRd + SMOTEd trained *SVM-RBF* vs. best of other classifiers including baseline models (Credit).

Classifier	N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
AdaBoostM1 vs *SVM-RBF*	14	267	33	686	23.40	[20.33, 26.47]	< 0.0001★
Random Forest vs *SVM-RBF*	9	219	38	734	18.10	[15.16, 21.04]	< 0.0001★
NB vs *SVM-RBF*	10	228	37	725	19.10	[16.14, 22.06]	< 0.0001★
RIPPER vs *SVM-RBF*	10	218	37	735	18.10	[15.18, 21.02]	< 0.0001★
C4.5 vs *SVM-RBF*	14	150	33	803	11.70	[9.15, 14.25]	< 0.0001★

N_{ff}: both models failed — N_{fs}: *C4.5* succeeded and the other classifier failed
N_{ss}: both models succeeded — N_{sf}: *C4.5* failed and the other classifier succeeded

Appendix B. QSAR Biodegradation data

The QSAR Biodegradation data from chemistry domain is used to classify chemicals into *ready* and *non-ready* biodegradable molecules. The data contains 1055 samples, each with 41 molecular descriptors as features. The *SMOTEd* version of the data was generated by simply oversampling the minority class by 96% using SMOTE. From the *original*, version, we identified 366 outlier instances using IQR algorithm. These instances were oversampled (without replacement) by 100% to generate additional 366 and subsequently added back to the data, i.e., a total of 732 outliers now exist. Finally, the minority class was oversampled by 100% using SMOTE to generate the *IQRd + SMOTEd* version of the data.

The results are presented in Tables B.1–B.3. We use **bold typeface** to indicate best performance among classifier groups, and '★' sign to indicate statistically significant difference.

Table B.1
Classifier performance with *original* and *preprocessed* versions of the *Biodegradation* dataset.

Classifier/data		Accuracy	Precision	Recall	F-Score	Kappa
Baseline	AdaBoostM1	0.811	0.813	0.811	0.812	0.711
	Random Forest	**0.857**	**0.857**	**0.857**	**0.857**	**0.783**
Naïve Bayes	Original	**0.759**	**0.825**	**0.759**	**0.766**	0.493
	SMOTEd	0.756	0.817	0.756	0.763	**0.498**
	IQRd + SMOTEd	0.743	0.815	0.743	0.750	0.447
SVM-RBF	Original	0.850	0.848	0.850	0.848	0.786
	SMOTEd	0.874	0.878	0.874	0.875	0.801
	IQRd + SMOTEd	**0.885**	**0.890**	**0.885**	**0.887**	**0.817**
RIPPER	Original	0.822	0.819	0.822	0.818	0.748
	SMOTEd	0.842	0.846	0.842	0.843	0.751
	IQRd + SMOTEd	**0.858**	**0.864**	**0.858**	**0.860**	**0.772**
C4.5	Original	0.824	0.824	0.824	0.824	0.734
	SMOTEd	0.853	0.857	0.853	0.854	**0.769**
	IQRd + SMOTEd	**0.856**	**0.862**	**0.856**	**0.858**	**0.769**

Table B.2
Mc Nemar's test showing performance differences between the *Biodegradation* data versions. Each dataset in the second column is compared against *IQRd + SMOTEd*.

Classifier/data		N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
NB	Original	81	173	190	611	−1.61	[−5.15, 1.93]	0.4011
	SMOTEd	247	10	24	774	−1.33	[−2.41, −0.25]	0.0243★
SVM	Original	22	136	99	798	3.51	[0.67, 6.35]	0.0187★
	SMOTEd	103	30	18	904	1.14	[−0.15, 2.42]	0.1114
RIPPER	Original	30	158	120	747	3.60	[0.51, 6.69]	0.0263★
	SMOTEd	82	85	68	820	1.61	[−0.68, 3.91]	0.1957
C4.5	Original	27	159	125	744	3.22	[0.098, 6.35]	0.0500
	SMOTEd	84	71	68	832	0.28	[−1.91, 2.47]	0.8654

N_{ff}: both models failed — N_{fs}: *IQRd + SMOTEd* succeeded and the other model failed
N_{ss}: both models succeeded — N_{sf}: *IQRd + SMOTEd* failed and the other model succeeded

Table B.3
IQRd + SMOTEd trained SVM-RBF vs. best of other classifiers including baseline models (biodegeneration dataset).

Classifier	N_{ff}	N_{fs}	N_{sf}	N_{ss}	*diff*	95% CI	*P*-value
AdaboostM1 vs *SVM-RBF*	21	178	1-00	756	7.39	[4.33, 10.46]	<0.00001★
Random Forest vs *SVM-RBF*	15	136	1-06	798	2.84	[−0.041, 5.73]	0.0621
NB vs *SVM-RBF*	31	223	90	711	12.61	[9.41, 15.80]	<0.0001★
RIPPER vs *SVM-RBF*	68	82	53	852	2.75	[0.60, 4.90]	0.0158★
C4.5 vs *SVM-RBF*	65	87	56	847	2.94	[0.72, 5.15]	0.0118★

N_{ff}: both models failed — N_{fs}: *C4.5* succeeded and the other classifier failed
N_{ss}: both models succeeded — N_{sf}: *C4.5* failed and the other classifier succeeded

References

[1] Zhai J, Zhang S, Wang C. The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. Int J Mach Learn Cybern 2017;8(3):1009–17. https://doi.org/10.1007/s13042-015-0478-7.
[2] Diabetes UK. Number of people with diabetes up 60 per cent in last decade. 2015 [accessed 29 July 2017]. https://www.diabetes.org.uk/About_us/News/diabetes-up-60-per-cent-in-last-decade-/.
[3] Czarnecki WM, Tabor J. Extreme entropy machines: robust information theoretic classification. Pattern Anal Appl 2017;20(2):383–400. https://doi.org/10.1007/s10044-015-0497-8.
[4] Skryjomski P, Krawczyk B. Influence of minority class instance types on smote imbalanced data oversampling. In: Torgo L, Krawczyk B, Branco P, Moniz N, editors. Proceedings of the first international workshop on learning with imbalanced domains: theory and applications (LIDTA 2017), Vol. 74 of proceedings of machine learning research, PMLR. Skopje, Macedonia: ECML-PKDD; 2017. p. 7–21http://proceedings.mlr.press/v74/skryjomski17a.html.
[5] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 2016;5(4):221–32. https://doi.org/10.1007/s13748-016-0094-0.
[6] Ferdowsi H, Jagannathan S, Zawodniok M. An online outlier identification and removal scheme for improving fault detection performance. IEEE Trans Neural Netw Learn Syst 2014;25(5):908–19. https://doi.org/10.1109/TNNLS.2013.2283456.
[7] Kaneda Y, Pei Y, Zhao Q, Liu Y. Improving the performance of the decision boundary making algorithm via outlier detection. J Inform Process 2015;23(4):497–504. https://doi.org/10.2197/ipsjjip.23.497.
[8] Dua D, Graff C. UCI machine learning repository. 2017http://archive.ics.uci.edu/ml.
[9] Upton G, Cook I. Understanding statistics. Oxford University Press; 1996.
[10] Cochran WG. Sampling techniques. 3rd Edition John Wiley; 1977.
[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority oversampling technique. J Artif Intell Res 2002;16(1):321–57http://dl.acm.org/citation.cfm?id=1622407.1622416.
[12] Diabetes UK. Diabetes: facts and stats. 2015 [accessed 6 May 2018]. https://www.mrc.ac.uk/documents/pdf/diabetes-uk-facts-and-stats-june-2015/.
[13] Roche MM, Wang PP. Factors associated with a diabetes diagnosis and late diabetes diagnosis for males and females. J Clin Transl Endocrinol 2014;1(3):77–84. https://doi.org/10.1016/j.jcte.2014.07.002.
[14] Holt TA, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. Br J Gen Pract 2008;58(548):192–6. https://doi.org/10.3399/bjgp08X277302.
[15] Holt TA, Gunnarsson CL, Cload PA, Ross SD. Identification of undiagnosed diabetes and quality of diabetes care in the united states: cross-sectional study of 11.5 million primary care electronic records. CMAJ Open 2014;2(4):E248–55. https://doi.org/10.9778/cmajo.20130095.
[16] Leong A, Dasgupta K, Chiasson J-L, Rahme E. Estimating the population prevalence of diagnosed and undiagnosed diabetes. Diabetes Care 2013;36(10):3002–8. https://doi.org/10.2337/dc12-2543.
[17] Bagheri N, McRae I, Konings P, Butler D, Douglas K, Del Fante P, et al. Undiagnosed diabetes from cross-sectional gp practice data: an approach to identify communities with high likelihood of undiagnosed diabetes. BMJ Open 2014;4:e005305. https://doi.org/10.1136/bmjopen-2014-005305.
[18] Sentell T, Cheng Y, Saito E, Seto T, Miyamura J, Mau M, et al. The burden of diagnosed and undiagnosed diabetes in native Hawaiian and Asian American hospitalized patients. J Clin Transl Endocrinol 2015;2(4):115–24. https://doi.org/10.1016/j.jcte.2015.08.002.
[19] Selvin E, Wang D, Lee A, Bergenstal RM, Coresh J. Identifying trends in undiagnosed diabetes in U.S. adults by using a confirmatory definition: a cross-sectional study. Ann Intern Med 2017;167(11):769–76. https://doi.org/10.7326/M17-1272.
[20] Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346(6):393–403. https://doi.org/10.1056/NEJMoa012512.
[21] Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, Brenneman AT, et al. 10-year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. Lancet 2009;374(9702):1677–86. https://doi.org/10.1016/S0140-6736(09)61457-4.
[22] The 10-year cost-effectiveness of lifestyle intervention or metformin for diabetes prevention. Diabetes Care 2012;35(4):723–30. https://doi.org/10.2337/dc11-1468.
[23] Rohwer R, Wynne-Jones M, Wysotzki F. Neural networks, Ellis Horwood series in artificial intelligence. Ellis Horwood; 1994. p. 84–106 The above book (originally published in 1994 by Ellis Horwood) is now out of print. The copyright now resides with the editors who have decided to make the material freely available on the web http://www1.maths.leeds.ac.uk/charles/statlog/.
[24] Quinlan JR. C4. 5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1993.
[25] John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. Proceedings of the eleventh conference on uncertainty in artificial intelligence, UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1995. p.

338–45http://dl.acm.org/citation.cfm?id=2074158.2074196.
[26] Cohen WW. Fast effective rule induction. Proceedings of the twelfth international conference on international conference on machine learning, ICML'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1995. p. 115–23http://dl.acm.org/citation.cfm?id=3091622.3091637.
[27] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
[28] Freund Y, Schapire RE. Experiments with a new boosting algorithm. Proceedings of the thirteenth international conference on international conference on machine learning, ICML'96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1996. p. 148–56http://dl.acm.org/citation.cfm?id=3091696.3091715.
[29] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
[30] Diabetes UK. Diabetes prevalence. 2017 [accessed 29 July 2017]. http://www.diabetes.co.uk/diabetes-prevalence.html.
[31] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16. https://doi.org/10.1016/j.csbj.2016.12.005.
[32] Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics 2016;17(1):359. https://doi.org/10.1186/s12859-016-1236-x.
[33] Carpenter GA, Markuzon N. Artmap-ic and medical diagnosis: Instance counting and inconsistent cases. Neural Netw: Off J Int Neural Netw Soc 1998;11(2):323–36.
[34] Deng D, Kasabov N. On-line pattern analysis by evolving self-organizing maps. Neurocomputing 2003;51:87–103. https://doi.org/10.1016/S0925-2312(02)00599-4.
[35] Kayaer K, Yildirim T. Medical diagnosis on pima indian diabetes using general regression neural networks. Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP) 2003:181–4.
[36] Winiarski T. Pima diabetes. 2003 [accessed 15 October 2018]. http://www.is.umk.pl/twin/pima_res.html.
[37] Polat K, Günes S, Arslan A. A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Syst Appl 2008;34(1):482–7. https://doi.org/10.1016/j.eswa.2006.09.012.
[38] Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl 2009;36(4):8610–5. https://doi.org/10.1016/j.eswa.2008.10.032.
[39] Ramezani R, Maadi M, Khatami SM. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Eng J 2017. https://doi.org/10.1016/j.aej.2017.03.043.
[40] Carpenter GA, GrossBerg S. Pattern Recognition by Self-organizing Neural Networks. Massachusetts: MIT Press; 1991.
[41] Carpenter GA, Grossberg S, Reynolds JH. ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Netw 1991;4(5):565–88. https://doi.org/10.1016/0893-6080(91)90012-T.
[42] Carpenter GA, GrossBerg S, Markuzon N, Reynolds JH, Rosen DB. Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans Neural Netw 1992;3(5):698–713.
[43] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. Smoteboost: improving prediction of the minority class in boosting. In: Lavrač N, Gamberger D, Todorovski L, Blockeel H, editors. Knowledge discovery in databases: PKDD 2003. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 107–19.
[44] Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Huang D-S, Zhang X-P, Huang G-B, editors. Advances in intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 878–87.
[45] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho T-B, editors. Advances in knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 475–82.
[46] Maciejewski T, Stefanowski J. Local neighbourhood extension of smote for mining imbalanced data. 2011 IEEE symposium on computational intelligence and data mining (CIDM) 2011:104–11. https://doi.org/10.1109/CIDM.2011.5949434.
[47] Barua S, Islam MM, Yao X, Murase K. Mwmote-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng 2014;26(2):405–25. https://doi.org/10.1109/TKDE.2012.232.
[48] Laurikkala J. Improving identification of difficult small classes by balancing class distribution. Proceedings of the 8th conference on AI in medicine in Europe: artificial intelligence medicine, AIME'01. Berlin, Heidelberg: Springer-Verlag; 2001. p. 63–6http://dl.acm.org/citation.cfm?id=648155.757340.
[49] Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B Cybern 2009;39(2):539–50. https://doi.org/10.1109/TSMCB.2008.2007853.
[50] García S, Herrera F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol Comput 2009;17(3):275–306. https://doi.org/10.1162/evco.2009.17.3.275.
[51] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C Appl Rev 2012;42(4):463–84. https://doi.org/10.1109/TSMCC.2011.2161285.
[52] Koziarski M, Wozniak M. Ccr: a combined cleaning and resampling algorithm for imbalanced data classification. Int J Appl Math Comput Sci 2017;27(4):727–36. https://doi.org/10.1515/amcs-2017-0050.
[53] Nanni L, Fantozzi C, Lazzarini N. Coupling different methods for overcoming the class imbalance problem. Neurocomputing 2015;158:48–61. https://doi.org/10.1016/j.neucom.2015.01.068.
[54] Raghuwanshi BS, Shukla S. Class imbalance learning using UnderBagging based kernelized extreme learning machine. Neurocomputing 2019;329:172–87. https://doi.org/10.1016/j.neucom.2018.10.056.
[55] Jegierski H, Saganowski S. An "outside the box" solution for imbalanced data classification. 2019arXiv:1911.06965.
[56] Stefanowski J, Krawiec K, Wrembel R. Exploring complex and big data. Int J Appl Math Comput Sci 2017;27(4):669–79. https://doi.org/10.1515/amcs-2017-0046.
[57] Napierała K, Stefanowski J, Wilk S. Learning from imbalanced data in presence of noisy and borderline examples rough sets and current trends in computing. Vol. 6086 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin /Heidelberg; 2010. p. 158–67. https://doi.org/10.1007/978-3-642-13529-3_18 [chapter 18].
[58] Deza MM, Deza E. Encyclopedia of distances. Springer Berlin Heidelberg; 2009. https://doi.org/10.1007/978-3-642-00234-2_1.
[59] Iba W, Langley P. Induction of one-level decision trees. Proceedings of the ninth international workshop on machine learning, ML'92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1992. p. 233–40http://dl.acm.org/citation.cfm?id=645525.757759.
[60] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–7. https://doi.org/10.1007/BF02294363.
[61] Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9(1):103. https://doi.org/10.1186/1741-7015-9-103.